Claims

[c1] 1. A method, comprising:

receiving one or more reservations for use of at least a first subset of a plurality of computing resources of a distributed computing system, wherein each of said one or more reservations specifies a period of time for use of said computing resources;

allocating said first subset of said computing resources for use in accordance with said one or more reservations;

receiving one or more requests for use of at least a second subset of said plurality of computing resources of said distributed computing system, wherein each of said one or more requests specifies a period of time for use of said computing resources;

determining whether a sufficient amount of one or more unallocated computing resources are available to fulfill all of said one or more requests, wherein said one or more unallocated computing resources comprises said computing resources of said distributed computing system that are not allocated in accordance with said one or more reservations;

responsive to said sufficient amount of said unallocated computing resources being available, allocating said unallocated computing resources in accordance with said one or more requests; and

responsive to said sufficient amount of said unallocated computing resources not being available, allocating said unallocated computing resources in accordance with an allocation criteria.

[c2] 2. The method of claim 1, further comprising modifying an amount of said plurality of computing resources of said distributed computing system based on consideration of said one or more reservations.

[c3] 3. The method of claim 2, wherein said modifying comprises adding a computing resource while said distributed computing system is in use.

[c4] 4. The method of claim 1, further comprising monitoring a usage level for at least a portion of said computing resources of said distributed computing system.

[c5] 5. The method of claim 4, further comprising providing data descriptive of said usage level.

[c6] 6. The method of claim 5, wherein said providing data descriptive of said usage level is performed in real time.

[c7] 7. The method of claim 5, further comprising using a graphical user interface to display said data descriptive of said usage level.

[c8] 8. The method of claim 7, wherein said usage level comprises a present usage of said plurality of computing devices.

[c9] 9. The method of claim 7, wherein said usage level comprises a historical usage of said plurality of computing devices.

[c10] 10. The method of claim 1, further comprising monitoring an allocation status for at least a portion of said computing resources of said distributed computing system.

[c11] 11. The method of claim 10, further comprising providing data descriptive of said allocation status.

[c12] 12. The method of claim 11, wherein said providing data descriptive of said allocation status is performed in real time.

[c13] 13. The method of claim 11, further comprising using a graphical user interface to display said data descriptive of said allocation status.

[c14] 14. The method of claim 13, wherein said allocation status comprises a present allocation of said plurality of computing resources.

[c15] 15. The method of claim 13, wherein said allocation status comprises a historical allocation of said plurality of computing resources.

[c16] 16. The method of claim 1, further comprising monitoring an inventory of computing resources of said distributed computing system.

[c17] 17. The method of claim 16, further comprising providing data descriptive of said inventory.

[c18] 18. The method of claim 17, wherein said providing data descriptive of said inventory is performed in real time.

[c19] 19. The method of claim 17, further comprising using a graphical user interface to display said data descriptive of said inventory.

[c20] 20. The method of claim 17, wherein said inventory comprises a present inventory of said plurality of computing resources.

[c21] 21. The method of claim 17, wherein said inventory comprises a historical inventory of said plurality of computing resources.

[c22] 22. The method of claim 1, wherein said plurality of computing resources comprises a processing device.

[c23] 23. The method of claim 1, wherein said plurality of computing resources comprises:
a first type of processing device having a first processing capability; and
a second type of processing device having a second processing capability, wherein said first and second processing capabilities are different.

[c24] 24. The method of claim 23, wherein each of said one or more reservations comprises an expression of said first and second types of processing device in a normalized unit of processing capability.

[c25] 25. The method of claim 23, wherein each of said one or more requests comprises an expression of said first and second types of processing device in a normalized unit of processing capability.

[c26] 26. The method of claim 1, further comprising charging a user for canceling a reservation.

[c27] 27. The method of claim 1, wherein said plurality of computing resources comprises a memory device.

[c28] 28. The method of claim 27, wherein said plurality of computing resources further comprises a processing device.

[c29] 29. The method of claim 1, further comprising billing a user of said computing resources.

[c30] 30. The method of claim 29, wherein said billing comprises determining whether a first price or a second price is to be billed.

[c31] 31. The method of claim 30, wherein said first price comprises a peak price.

[c32] 32. The method of claim 30, wherein said first price comprises an off-peak price.

[c33] 33. The method of claim 32, wherein said second price comprises a peak price.

[c34] 34. The method of claim 30, wherein said first price is billed for said computing resources allocated in response to said reservation and said second price is billed for computing resources allocated in response to said request, wherein said first price is higher than said second price.

[c35] 35. The method of claim 1, wherein said one or more requests each comprise a priority indication, and wherein said allocation criteria considers said priority indication of each said request.

[c36] 36. The method of claim 35, wherein said allocation criteria comprises a calculation of a weighted average based at least in part on said priority indications.

[c37] 37. The method of claim 35, further comprising billing a user of said computing resources such that a cost varies in accordance with said priority indication.

[c38] 38. The method of claim 1, wherein said allocation criteria comprises an equal division of said unallocated computing resources between a plurality of users that have made a request.

[c39] 39. The method of claim 1, wherein said one or more requests each comprise a bid indication, and wherein said allocation criteria considers said bid indication of each said request.

[c40] 40. The method of claim 39, wherein said allocation criteria comprises fulfilling

said requests beginning with said request comprising a highest bid indication and continuing in descending order of requests comprising said bid indications of lesser values until all of said unallocated resources have been allocated.

[c41] 41. The method of claim 1, wherein said unallocated computing resources are allocated dynamically.

[c42] 42. The method of claim 41, further comprising re-allocating said unallocated computing resources dynamically.

[c43] 43. The method of claim 1, wherein said unallocated computing resources are allocated in real time in response to receiving said one or more requests.

[c44] 44. A system, comprising:

means for receiving one or more reservations for use of at least a first subset of a plurality of computing resources of a distributed computing system, wherein each of said one or more reservations specifies a period of time for use of said computing resources;

means for allocating said first subset of said computing resources for use in accordance with said one or more reservations;

means for receiving one or more requests for use of at least a second subset of said plurality of computing resources of said distributed computing system, wherein each of said one or more requests specifies a period of time for use of said computing resources;

means for determining whether a sufficient amount of one or more unallocated computing resources are available to fulfill all of said one or more requests, wherein said one or more unallocated computing resources comprises said computing resources of said distributed computing system that are not allocated in accordance with said one or more reservations;

means for allocating said unallocated computing resources in accordance with said one or more requests and in response to said sufficient amount of said unallocated computing resources being available; and

means for allocating said unallocated computing resources in accordance with an allocation criteria and in response to said sufficient amount of said unallocated computing resources not being available.

[c45] 45. A system, comprising:

a distributed computing system comprising a plurality of computing resources; and

a computing device configured to:

receive one or more reservations for use of at least a first subset of said plurality of computing resources, wherein each of said one or more reservations specifies a period of time for use of said computing resources;

allocate said first subset of said computing resources for use in accordance with said one or more reservations;

receive one or more requests for use of at least a second subset of said plurality of computing resources of said distributed computing system, wherein each of said one or more requests specifies a period of time for use of said computing resources;

determine whether a sufficient amount of one or more unallocated computing resources are available to fulfill all of said one or more requests, wherein said one or more unallocated computing resources comprises said computing resources of said distributed computing system that are not allocated in accordance with said one or more reservations;

responsive to said sufficient amount of said unallocated computing resources being available, allocate said unallocated computing resources in accordance with said one or more requests; and

responsive to said sufficient amount of said unallocated computing resources not being available, allocate said unallocated computing resources in accordance with an allocation criteria.

[c46] 46. The system of claim 45, wherein said computing device comprises a server.

[c47] 47. The system of claim 45, wherein said computing device comprises at least two servers, wherein each server is in a different geographic location.

[c48] 48. The system of claim 45, further comprising a graphical user interface configured to display data descriptive of a usage level for at least a portion of

said plurality of computing resources of said distributed computing system.

[c49] 49. The system of claim 48, wherein said computing device is further configured to provide data descriptive of said usage level in real time.

[c50] 50. The system of claim 45, further comprising a graphical user interface configured to display data descriptive of an allocation status for at least a portion of said computing resources of said distributed computing system.

[c51] 51. The system of claim 50, wherein said computing device is further configured to provide data descriptive of said allocation status in real time.

[c52] 52. The system of claim 45, further comprising a graphical user interface configured to display data descriptive of an inventory of computing resources of said distributed computing system.

[c53] 53. The system of claim 52, wherein said computing device is further configured to provide data descriptive of said inventory in real time.

[c54] 54. The system of claim 45, wherein said computing device is further configured to generate a billing record based on a usage level of said plurality of computing resource.

[c55] 55. The system of claim 45, further comprising a persistent data storage queue in communication with said computing resources, and wherein a minimum availability of said distributed computing system is defined by an availability of said persistent data storage queue.

[c56] 56. The system of claim 45, wherein said plurality of computing resources comprises a processing device.

[c57] 57. The system of claim 45, wherein said plurality of computing resources comprises:
a first type of processing device having a first processing capability; and
a second type of processing device having a second processing capability, and wherein said first and second processing capabilities are different.

[c58] 58. The system of claim 57, wherein each of said one or more reservations

comprises an expression of said first and second types of processing device in a normalized unit of processing capability.

[c59] 59. The system of claim 57, wherein each of said one or more requests comprises an expression of said first and second types of processing device in a normalized unit of processing capability.

[c60] 60. The system of claim 45, wherein said computing device is further configured to charge a user for canceling a reservation.

[c61] 61. The system of claim 45, wherein said plurality of computing resources comprises a memory device.

[c62] 62. The system of claim 61, wherein said plurality of computing resources further comprises a processing device.

[c63] 63. The system of claim 45, wherein said allocation criteria considers said one or more priority indications of said requests.

[c64] 64. The system of claim 45, wherein said computing device is further configured to bill a user of said computing resources.

[c65] 65. The system of claim 64, wherein said computing device is further configured to determine whether a first price or a second price is to be billed.

[c66] 66. The system of claim 65, wherein said first price comprises a peak price.

[c67] 67. The system of claim 65, wherein said first price comprises an off-peak price.

[c68] 68. The system of claim 67, wherein said second price comprises a peak price.

[c69] 69. The system of claim 65, wherein said first price is billed for said computing resources allocated in response to said reservation and said second price is billed for computing resources allocated in response to said request, wherein said first price is higher than said second price.

[c70] 70. The system of claim 45, wherein said one or more requests each comprise a priority indication, and wherein said allocation criteria considers said priority

indication of each said request.

[c71] 71. The system of claim 70, wherein said allocation criteria comprises a calculation of a weighted average based at least in part on said priority indications.

[c72] 72. The system of claim 70, wherein said computing device is further configured to bill a user of said computing resources such that a cost varies in accordance with said priority indication.

[c73] 73. The system of claim 45, wherein said allocation criteria comprises an equal division of said unallocated computing resources between a plurality of users that have made a request.

[c74] 74. The system of claim 45, wherein said one or more requests each comprise a bid indication, and wherein said allocation criteria considers said one or more bid indication of each said request.

[c75] 75. The system of claim 74, wherein said allocation criteria comprises fulfilling said requests beginning with said request comprising a highest bid indication and continuing in descending order of requests comprising said bid indications of lesser values until all of said unallocated resources have been allocated.

[c76] 76. The system of claim 45, wherein said computing device is further configured to dynamically allocate said unallocated computing resources.

[c77] 77. The system of claim 45, wherein said computing device is further configured to allocate said unallocated computing resources in real time in response to receiving said one or more requests.